DOCUMENT RESUME

ED 318 574 PS 018 762

AUTHOR Horm-Wingerd, Diane M.
TITLE Assesssing Tests: What Do Teachers Need To Know?
PUB DATE Nov 88
NOTE 21p.; A portion of this paper was presented at the Annual Meeting of the National Association for the Education of Young Children (Anaheim, CA, November 10-13, 1988).
PUB TYPE Guides - Non-Classroom Use (055) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Educational Practices; Guidelines; Norm Referenced Tests; Resource Materials; Standardized Tests; *Teacher Responsibility; Test Reliability; *Test Selection; *Test Use; Test Validity

ABSTRACT

This paper provides an overview of standardized tests and a summary of important psychometric and test construction properties that educators must consider when they make critical evaluations of standardized tests. Sections of the paper concern: (1) basic characteristics of standardized tests; (2) differences between norm-referenced and criterion-referenced tests; (3) reliability and validity; (4) the importance of reliability and validity; and (5) sources of additional information. Guidelines for appropriate test selection and use are included. (Author/RH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



ED318574

# Assessing Tests: What Do Teachers Need To Know?

Diane M. Horm-Wingerd

Department of Human Development, Counseling, and Family Studies

The University of Rhode Island

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Diane M. Horm-Wingerd

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running head: ASSESSING TESTS

A portion of this paper was presented at the Annual Meeting of the National Association for the Education of Young Children held in Anaheim, CA, November 1988.

For further information contact: Diane M. Horm-Wingerd, Assistant Professor, Department of Human Development, Counseling, and Family Studies, Transition Center, The University of Rhode Island, Kingston, RI 02881. Departmental phone number is (401) 792-2150.

PS 018762

Abstract

The testing of young children has become quite pervasive. In today's world, teachers are often responsible for choosing tests, administering tests, or using test results. Teachers, many who have not received formal training in testing, report feeling overwhelmed and frustrated with this relatively new responsibility. The purpose of this article is to help teachers become familiar with some important concepts related to standardized testing. The article provides an overview of standardized tests and summarizes the important psychometric and test construction properties that must be considered when critically evaluating standardized tests. Guidelines for appropriate test selection and use are included.

Assessing Tests: What Do Teachers Need To Know?

"I don't care what the statistics show, I know that test produces helpful results." "We do not have the luxury of time to thoroughly review the properties of tests--our kids need programs." Both of these comments were made by dedicated, caring educators during discussions of various testing programs and practices. Although the comments were motivated by deep concern for children, the comments show these educators were unfamiliar with or undervalued the importance of psychometric data when initially selecting and choosing to continue a testing program.

These educators are not alone. In today's world, many teachers are responsible for choosing tests, administering tests, or using test results. The testing of young children has become quite pervasive in contemporary society, especially in regard to placement decisions for 4-, 5-, and 6-year-olds (National Association for the Education of Young Children [NAEYC], 1988). However, most teachers have not received formal training in testing and feel overwhelmed with this relatively new responsibility. The purpose of this article is to help teachers become more familiar with some important concepts related to standardized testing. Although there is much to learn about testing, some of the most important information about standardized tests is related to test construction and the psychometric properties of reliability and validity. Thus, this article will provide an overview of standardized tests and summarize the important psychometric properties that must be considered when evaluating standardized tests. Guidelines for educators to employ when selecting and using tests will also be offered.

What are Standardized Tests? Standardized tests can be defined as tests which include fixed administration and scoring procedures, standard apparatus and format, and empirical testing of items. Once children are exposed to the resulting common task, their performance is interpreted relative to some set standard (Goodwin & Driscoll, 1980). Most standardized tests are commercially published and intended for widespread use. Thus, the content tends to be broad and general and not specifically related to the objectives of an individual classroom teacher (Goodwin & Driscoll, 1980).

It is often easier to understand standardized tests when they are contrasted with non-standardized tests, the most common example of the latter being teacher-constructed tests. Most teacher-constructed tests, as the name implies, are constructed by teachers for use in their classrooms for planning or evaluative purposes. These tests are typically administered only once to a particular class and cover a defined and specific set of material which is closely related to the curriculum (Brown, 1976). Thus, standardized and non-standardized tests differ in content, purpose, and use (Goodwin & Driscoll, 1980).

Norm-referenced vs. criterion-referenced tests. Performance on standardized tests may be interpreted using either norm-referenced or criterion-referenced standards (Brown, 1976). The choice depends on the goals of testing or on the type of information that is desired (Goodwin & Driscoll, 1980).

With a norm-referenced test, a child's performance is compared with a norm group or a group of other children who took the same test--hence the

label norm-referenced. For example, a norm-referenced reading test would indicate how Mark performed compared to other second grades who took the test. Scores on norm-referenced tests are reported as relative standing within a group (Brown, 1976). So, if Mark's performance fell at the 50th percentile, the teacher would know he was demonstrating average reading skills as compared to other second graders.

With criterion-referenced tests, a child's performance is compared against predetermined levels of mastery or an established standard of performance. A child is compared to some established criterion rather than to other children--hence the label criterion-referenced (Sattler, 1982). For example, a criterion of 10 math skills could be determined for first graders. Such skills could include matching, sorting, categorizing, 1-1 correspondence, seriation, etc. A criterion-referenced test would indicate to Ms. Groves which children in her class had mastered each particular skill. Thus, criterion-referenced tests usually categorize children into two groups--those who have mastered the material and those who have not (Goodwin & Driscoll, 1980). As can be seen, criterion-referenced tests provide information which is more closely related to instructional decisions than norm-referenced tests. Such information includes what skills a child demonstrates, if there are certain skills that require more attention, etc. (Sattler, 1982).

Norm-referenced and criterion-referenced testing play different roles and make different contributions to our understanding of children (Sattler, 1982). Norm-referenced testing enables comparisons of an individual child's performance to the performance of a group of children

(Goodwin & Driscoll, 1980). Criterion-referenced tests show whether or not a child can perform certain behaviors (Goodwin & Driscoll, 1980). These two types of testing are essentially complementary methods of assessment, with each telling us different types of information. Although, criterion-referenced tests are becoming more popular, norm-referenced tests are more commonly used (Goodwin & Driscoll, 1980; Patton, 1975). Thus, a consideration of norms is important.

Norms

Norms provide information about how children from some comparison group actually performed on a test. Thus, norms contain information about a group of children's actual performance on a test, not how children ought to have performed (Patton, 1975). Norms can be reported in several ways such as standard scores, percentile ranks, grade equivalents, and stanines (Patton, 1975). In essence, norms are a numerical description of the test performance of a reference group of children (Goodwin & Driscoll, 1980).

Because norms are the standard to which an individual child's performance is compared, a close inspection of the norms is essential in determining a test's usefulness (Goodwin & Driscoll, 1980). Norms should have certain characteristics to ensure proper interpretation. A list of some important characteristics includes:

1. Representativeness. Representativeness of the norm group refers to the extent to which the group is characteristic of a particular population (Sattler, 1982). If a test is to be used nationally with children from different races, socioeconomics levels, geographical locations, etc., then the normative sample should include children

representing all levels of these and other important demographic variables (Goodwin & Driscoll, 1980). Generally, the most important demographic characteristics are age, grade level, sex, ethnicity, geographic region, and socioeconomic status (Sattler, 1982). Norms are limited if they are developed on a group of similar children from one specific region of the country (Goodwin & Driscoll, 1980). Such limited norms can lead to serious misinterpretations. In fact, the representativeness of the normative group is more important than its size (Goodwin & Driscoll, 1980).

2. Subgroup norms. National norms which should match the various demographic characteristics of the population as a whole may not always be the most appropriate. In some cases, norms of a particular segment of the population may be the most relevant (Sattler, 1982). For this reason, test publishers should also include subgroup norms according to sex, race, region, etc. This is an increasingly common practice and facilitates the proper interpretation of results (Goodwin & Driscoll, 1980). Selection of the correct norm group against which to evaluate a child's test performance is critically important for proper test interpretation (Sattler, 1982). As stated by Sattler, "The person interpreting a test must possess the skills and insight to select the proper norm group for each individual being evaluated" (p. 17). This insight stems from the vast knowledge teachers have about the children and families they serve. Any testing report should clearly state which normative group has been selected. This is especially true if the norm group is different than the one commonly used (Sattler, 1982).

3. Size of normative sample. Although representativeness is important, norms based on larger samples are more stable (Sattler, 1982). Sattler recommends that if a test includes different age or grade levels that at least 100 children from each age or grade level be included in the normative group.

Psychometric Characteristics of Reliability and Validity

In addition to assessing the adequacy of the norms, a test user must also ask some other important questions about tests. Such questions include: How consistent are scores on the test? What does the test really measure? These questions relate to reliability and validity. While the choice of a particular type of test (norm-referenced vs. criterion-referenced; cognitive vs. emotional domain) depends on the purpose of the assessment, the specific test ultimately employed should be selected on the basis of reliability and validity (Goodwin & Driscoll, 1980).

Reliability. "Reliability refers to the consistency of scores obtained by the same person when reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions" (Anastasi, 1982, p. 102). As suggested, consistency can be defined in different ways and there are various ways to estimate the consistency of a test. Three types of reliability are commonly calculated: Test-retest, alternate form, and internal consistency (Sattler, 1982).

Test-retest reliability is the most common procedure used to estimate the consistency of a test (Sattler, 1982). With test-retest reliability the same measure is administered twice to a single group of children

(Goodwin & Driscoll, 1980). The relationship between the two sets of scores represents the degree to which the test is consistent or stable over time. The higher the reliability, the less the test scores are influenced by the random changes in the conditions of the examinee or the testing environment on the two occasions.

With test-retest reliability, shorter retest intervals generally yield higher reliability estimates (Sattler, 1982). Thus, the interval between retests should always be specified in a test manual (Anastasi, 1982). As noted by Anastasi, the interval between tests should rarely exceed six months. This period should be shorter for tests designed for young children because developmental change is noticeable over a period of a month or less. Thus, with tests designed for young children, test-retest intervals should rarely exceed one month.

Although this technique is simple and straightforward, it has several problems. Problems include the fact that practice gained on the first administration of the test might affect retest scores and thus influence the estimate of stability (Anastasi, 1982). A second problem is that estimates can be influenced by memory--children may respond to the retest based on their memory of their responses on the first test. Thus, memory alone can account for high test-retest estimates (Anastasi, 1982). Because of these problems, test-retest procedures are not the best or most appropriate method to estimate the consistency or reliability of a test.

Alternate form reliability, often referred to as equivalent or parallel form reliability, is assessed by administering two equivalent tests to the same children at the same time (Sattler, 1982). The

relationship between the two sets of scores is an indicator of the consistency of the tests in regard to the content covered by the tests (Goodwin & Driscoll, 1980). Thus, this type of reliability assesses consistency in performance from one set of items to another, not consistency with the same set of items over time (Anastasi, 1982). Alternate form reliability also has certain limitations, however, it is generally more applicable than test-retest procedures (Anastasi, 1982). The reader is referred to Anastasi for a thorough discussion of weaknesses associated with reliability estimates derived from alternate form procedures.

Internal consistency is calculated in a variety of ways (split-half, Kuder-Richardson, coefficient alpha) but the various methods all yield information related to the homogeneity or the degree of consistency of the content within a single test form (Goodwin & Driscoll, 1980). High estimates of internal consistency suggest that all of the items included on a test measure the same trait or content area (Sattler, 1982). Temporal stability is not assessed because only one test session is involved (Anastasi, 1982).

The three forms of reliability discussed above have strengths and weaknesses that complement one another. Thus, ideally, more than one type of reliability should be calculated and reported for a test. It should be noted that no test is a perfectly reliable instrument (Anastasi, 1982). Every test should report estimates of its reliability and this information should be reported in the test manual or technical report which accompanies the test (American Educational Research Association, American

Psychological Association, and National Council on Measurement in Education [AERA, APA, NCME], 1985). Theoretically, reliability coefficients can range from 1.0, or perfect reliability to 0.0, or a complete lack of reliability. The closer the estimate is to 1.00, the higher the consistency of the scores. A measure with high reliability is considered accurate in the sense of producing stable or consistent scores.

Teachers should review the test manuals and related technical reports to evaluate the reliability of a test. Below are some guidelines suggested by Lehr, Ysseldyke, and Thurlow (1987):

1. At least one type of reliability should be reported. As suggested above, more than one type should ideally be reported.

2. Reliability procedures and samples should be reported and described. As indicated, the various procedures have different strengths and weaknesses so to be informed the test user must know how the reliability estimate was calculated. The characteristics of the reliability sample are also important for test evaluation. A reported reliability coefficient is applicable only to samples that are similar to the one used for the calculation of the reliability estimate.

3. Reliability coefficients of .90 or greater are necessary when tests are being used to make decisions about individuals (e.g., placements, diagnostic testing).

4. Reliability coefficients of .80 or greater are necessary when tests are being used for screening purposes (to identify children who may require further in-depth testing).

5. Reliability coefficients of .60 or greater are acceptable when group tests are given for administrative purposes (e.g., evaluate schools).

Validity. "The validity of a test concerns what the test measures and how well it does so" (Anastasi, 1982, p. 131). Procedures for determining test validity basically assess the relationships between test performance and other indices of the behavior in question (e.g., intelligence, math achievement, etc.) (Anastasi, 1982). As with reliability, different methods are used to determine test validity. The three principle types of validity are content, criterion-related, and construct validity (AERA, APA, NCME, 1985). Thorough validation of a test often involves all three types of validity. However, all three types are distinct and require different evidence (Goodwin & Driscoll, 1980).

Content validity is the extent to which the items on a test are a representative sample of the larger domain in question (Anastasi, 1982; Goodwin & Driscoll, 1980). Content validity is a matter of judgment. Content validity is typically determined by subject-matter experts or test users through logical analysis (Goodwin & Driscoll, 1980). According to Sattler (1982), when evaluating content validity one should "consider the appropriateness of the type of items, the completeness of the item sample, and the way in which the items assess the content of the domain involved" (p. 23). When evaluating a test for local use, it is suggested the test be examined item by item in light of local goals, programs, and curriculum (Goodwin & Driscoll, 1980).

Criterion-related validity assesses the relationship between test scores and a criterion or a direct and independent measure of the same

behavior the test claims to measure (Anastasi, 1982). Typical criteria are classroom performance, teacher ratings, or other test scores. There are two types of criterion-related validity: concurrent and predictive (Sattler, 1982). These two differ in terms of when the criterion measure is obtained. If the criterion is obtained at approximately the same time as the test score, the relationship between the two is a measure of concurrent validity. When there is a time interval between the test administration and performance on the criterion measure, the relationship between the two is a measure of predictive validity (Anastasi, 1982). Thus, concurrent validity addresses the question if a test can accurately replace a currently used procedure (or criterion). Predictive validity addresses a totally different question--is the test score an accurate predictor of future performance on the criterion? (Sattler, 1982). Predictive validity is very important given a frequent purpose of educational testing which is to predict future performance in school.

Construct validity is the extent to which a test measures a theoretical construct or trait (Sattler, 1982). Examples of constructs include intelligence, verbal fluency, and creativity (Anastasi, 1982). Construct validity is determined by the gradual accumulation of data from a variety of empirical studies (Goodwin & Driscoll, 1980). Based on theory, predictions are made about how performance on the test should relate to other variables. Construct validity is demonstrated when tests are related to variables with which they theoretically should correlate and do not relate to variables with which they theoretically should differ (Goodwin & Driscoll, 1980).

The validity of a test cannot be reported in general terms such as "high" or "low" (Anastasi, 1982). To quote Sattler (1982),"Tests are valid only for a specific purpose" (p. 23). No one single type of validity is appropriate for all testing purposes. A measure may be valid for one purpose but not for others (Goodwin & Driscoll, 1980). Thus, the question of validity is always tied to specific uses. As noted by Goodwin and Driscoll, the important question is not "Is this test valid?" but rather "Is this test valid for this particular purpose?" (p. 77).

Teachers need to carefully evaluate the question of test validity. Following are some of the guidelines provided by AERA, APA, and NCME (1985) for use in evaluating test validity. The reader is referred to their Standards for Educational and Psychological Testing for a more detailed consideration of requirements for validity.

1. Evidence for and statements about validity should address the validity of particular interpretations or of specific types of decisions. As indicated above, no test is valid for all purposes or in all situations. Thus, evidence of validity must be presented to support the recommended uses of a test. For example, if a test is used to predict future academic performance then it must demonstrate adequate predictive validity.

2. The samples used in validation studies should be clearly described so the test user can judge the pertinence to his or her situation. Relative to educational settings, important information about the sample includes gender, ethnic composition, characteristics of the

community, etc. The sample employed in the validation studies should mirror the characteristics of the persons for whom the test is recommended.

Importance of reliability and validity. Teachers often believe that reliability and validity are "just statistics" that have little relevance. However, tests that do not demonstrate acceptable reliability and validity are "inherently untrustworthy and should not be used to identify and place children" (Meisels, 1987, p. 5). When the psychometric criteria of reliability and validity are inadequately met or not reported, teachers risk misusing and abusing tests. Such misuse and abuse can have serious consequences. For example, children who do and do not require special services may be misidentified and thus not receive appropriate educational programs (Meisels, 1985).

Despite this knowledge, educators persist in using unreliable and invalid tests (Meisels, 1987). For example, a study conducted by Lehr et al. (1987) which examined current assessment practices used in model special education programs throughout the nation found that the vast majority of tests used were technically inadequate. Specifically, these researchers found that out of 19 frequently used tests, only three had technically adequate norms, reliability and validity. This was the case even though the respondents reported that they initially selected and continued to use the tests largely on their technical merits. Lehr et al. (1987) hypothesized that practitioners might attempt to explain the discrepancy between the reported influences of technical adequacy and the actual psychometric properties of the test by stating that although certain tests are inadequate, they are the only tests available. However,

as emphasized by Lehr et al. (1987), "using a technically inadequate device cannot be justified or excused" (p. 397). This is true because these tests are often used to make critically important decisions about children.

Some Reminders, Cautions, and Sources of Additional Information:

The fact that a test is published does not ensure quality (Goodwin & Driscoll, 1980). To critically evaluate tests, the test user must know what evidence to look for, how to evaluate that evidence, and how to initially acquire the necessary information or evidence (Goodwin & Driscoll, 1980). This article has highlighted what evidence to look for and how to evaluate that evidence. However, the issue of how to acquire the necessary evidence has not yet been addressed.

For test reviews which are completed by independent evaluators not involved in the construction of a particular test see Buros' Mental Measurement Yearbooks (MMY). The MMY provide facts, critical reviews, and bibliographies of tests published in English-speaking countries.

Another useful reference book is Buros' Tests in Print II (TIP II). The purpose of TIP II "is to direct readers to the MMY test reviews, the excerpted test reviews from journals, and to the professional literature on the construction, use, and validity of the tests being considered" (Buros, 1974, p. xxii).

To obtain a suggested outline helpful in conducting a thorough and systematic evaluation of tests, the reader is referred to Appendix C (p. 666-667) of Anastasi's (1982) book entitled Psychological Testing. This

outline contains the basic features which should be considered when critically evaluating a test.

For more information on psychometric considerations, standards for test use, and standards for test manuals, the reader is referred to Standards for Educational and Psychological Testing. The book contains a series of standards developed by a joint committee of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education which was most recently revised in 1985. The books mentioned above can be found in most college or university libraries.

The fact that a test is published also does not ensure it will be useful in improving services or outcomes for children. Educators must recognize that standardized tests have limitations and that even the best ones are not easy solutions for the problems faced by educators. In fact, the misuse and over-reliance on tests is the source of some common problems such as teaching to the test which fosters an inappropriate curriculum. For a discussion of some of the problems associated with tests and testing in early childhood the reader is referred to the "NAEYC position statement on standardized testing of young children 3 through 8 years of age" (NAEYC, 1988).

## Conclusion

In sum, it is important to remember that standardized tests are tools and like any tool their effectiveness depends upon the skill and expertise of the user (Sattler, 1982). When used cautiously and wisely, tests can help us obtain valuable insights which may facilitate our work with

children. However, when used inappropriately, tests may mislead and cause damaging results (Sattler, 1982). As professional educators, we must never forget this warning. We must also take steps to upgrade our competencies as test users so we avoid misusing and over-using standardized tests. "Tests will improve in their construction and uses only when the examiners and test consumers become confident in their ability to evaluate tests critically and to say 'NO' to the poor ones" (Weiner & Stewart, 1984, p. 2). Such informed and responsible decisions by teachers and others will pressure commercial test developers to produce and market valid, reliable, and useful tests (NAEYC, 1988). It is hoped this article serves to help teachers and others concerned with children take the first steps in becoming competent test users and consumers.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA & NCME]. (1985). Standards for educational and psychological testing. Washington, DC: Authors.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Brown, F. G. (1976). Principles of educational and psychological testing (2nd ed.). New York: Holt, Reinhart & Winston.

Buros, O. K. (1974). Tests in print II. Highland Park, NJ: Gryphon Press.

Buros, O. K. (1978). The mental measurement yearbook. Highland Park, NJ: Gryphon Press.

Goodwin, W., & Driscoll, L. (1980). Handbook for measurement and evaluation in early childhood education. San Francisco, CA: Jossey-Bass.

Lehr, C. A., Ysseldyke, J. E., & Thurlow, M. L. (1987). Assessment practices in model early childhood special education programs. Psychology in the Schools, 24, 390-398.

Meisels, S. J. (1985). Developmental screening in early childhood: A guide. Washington, DC: NAEYC.

Meisels, S. J. (1987). Uses and abuses of developmental screening and school readiness testing. Young Children, 42(2), 4-6; 68-73.

NAEYC. (1988). NAEYC position statement on standardized testing of young children 3 through 8 years of age. Young Children, 43(3), 42-47.

Patton, M. Q. (1975). Understanding the gobble-dy-gook. In V. Perrone, M. D. Cohen, and L. P. Martin (Eds.), Testing and evaluation: New views (pp. 18-26). Washington, DC: Association for Childhood Education International.

Sattler, J. M. (1982). Assessment of children's intelligence and special abilities (2nd ed.). Boston, MA: Allyn & Bacon.

Weiner, E.A., & Stewart, B. J. (1984). Assessing individuals: Psychological and educational tests and measurements. Boston, MA: Little, Brown & Co.